Research of the Bank's CRM Based on Data Mining Technology

Yong Wang, Dong Sheng Wu
School of Economics and Business Administration, Chongqing University
Chongqing China

***Abstract*-Customers are the base of existence and development of the banks. Good CRM (customer relationship management) can help the banks to find potential new customers and keep good relationships with old customers, so it is very important to the banks which are in the intense competitive environment. With the rapid development of the data mining technology, its effect in the bank's CRM will be more and more outstanding. Various data mining technologies can be comprehensively used not only to find valuable knowledge and laws in the sea of information in the bank's CRM, but also to explore the potential high-value relationships of the customers and forecast the behavior of the customers. This paper expounds the composition and major function of the bank's CRM, and constructs decision tree to analyze the kind of the bank's customers by applying the ID3 algorithm. This will attain the intellectual need in the CRM interactive process, help the bank understand the behavior of the customers to a fuller extent, and improve the service level of the bank.**

Keywords- the bank's CRM; data mining; ID3 algorithm

I. DEFINITION OF DATA MINING AND CATEGORIES OF ANALYSIS METHODS

A. Definition of Data Mining

With the development of computers and network technology, the capacities in the aspects of information production and data collection have been dramatically improved. A lot of databases have been widely applied in the fields of industrial production, business management, scientific research as well as administration of general government affairs. In this case, data explosion caused by data overload may bring many new problems. Therefore, people are in great need of related methods and technologies to automatically and intelligently acquire useful knowledge and information out of mass data. The formation and development of Data Mining are mainly based on the needs for the analysis and comprehension of data.

Data Mining can also be called Data Digging or Data Exploiting. A generally accepted definition of Data Mining is brought forward by W．J．Frawley and G．Piatetsk Shap iro: Data Mining is the extraction of interesting patterns or knowledge from huge amount of data systems or databases; or in other words, abstracting the subtle relationship that cannot be easily perceived out of the mass observation data, and finally drawing out a useful and understandable conclusion. In brief, Data Mining is to find out the relationships among patterns, knowledge and data.

The main analysis methods of Data Mining include Classification Analysis, Association Rule Analysis, Cluster Analysis and Sequential Pattern Analysis, etc.

B. Main Analysis Methods of Data Mining

1) Classification Analysis

Classification Analysis is defined as making accurate description of each category, setting up classification models or digging out the classification rules through data analysis in the database; then classifying the other records in accordance with such classification rules.

2) Association Rule Analysis

Association Rule Analysis is mainly to find out the relationship between different events: one event usually occurs with another one. The focus of Association Rule Analysis is on finding out the correlated events with practical value and the finding builds on the evidence that the probability or the conditional probability of an event's occurring is in line with the statistical significance.

3) Cluster Analysis

Cluster Analysis is to assign the record clusters reasonably according to a certain classification rule through the analysis of the recorded data in the database, and then putting the similar ones into one data cluster.

4) Sequential Pattern Analysis

Sequential Pattern Analysis is to dig out the relation schema of time sequence between different events. It emphasizes particularly on analyzing the sequential relationship among data and discovering internal transaction pattern of "some items follow the others" in the chronological transaction set.

II. COMPOSITION OF THE BANK'S CRM

A. Introduction of CRM

CRM is the abbreviation of Customer Relationship Management which means managing a company's interactions with customers. CRM is a newly-developed management mechanism aiming at enhancing the relationship between companies and customers. Meanwhile, it is also a company-wide business strategy which involves the whole process of using techniques of modern management and information to estimate, select, win and retain customers. Companies implementing the CRM systems are usually focused on customer relationship and are willing to ensure more satisfactory customer experience through reconstructing organizational structures, optimizing business processes and carrying out systematic researches on the customers. Therefore, customer values to the companies are greatly increased and thus increasing the companies' efficiencies and profits.

B. Composition of the Bank's CRM

The bank's CRM system is usually composed of Business Operation Subsystem, Customer Cooperation Subsystem, Data Analysis Management Subsystem and Integration of Bank Application Systems.

Business Operation Subsystem is mainly to ensure that the best methods are adopted to achieve the best result in the bank's sales and service processes through the establishment and management of business processes such as marketing and service processes with the support of computers and network technologies. It provides a processing platform for various products and services to the customers and records all data and information generated in the business activities into the database.

The main purpose of Customer Cooperation Subsystem is to provide the administrative staff and customer service staff with information required and diversified customer service channels. It includes Contact Management, Time Management, Call Center Management, Customer Service Center Management, Potential Customer Management and Web/E-mail Management.

Data Analysis Management Subsystem is the core of CRM system which is based on the central database of the system. Information is collected by various channels such as Customer Cooperation Subsystem and Business Operation Subsystem when contacting with customers and providing services for them. Then the data is summarized and classified. With the application of On-Line Analysis Processing (OLAP), Data Mining and other intelligence technologies, it can help the banks to fully understand the classifications, behaviors, needs and satisfaction of customers to seek for potential markets and forecast possible risks.

Integration of Bank Application Systems may be combined with the other subsystems to provide necessary interfaces between the subsystems. Thus, data contained in different systems can be completely shared which ensures the normal operation of Data Analysis Management Subsystem.

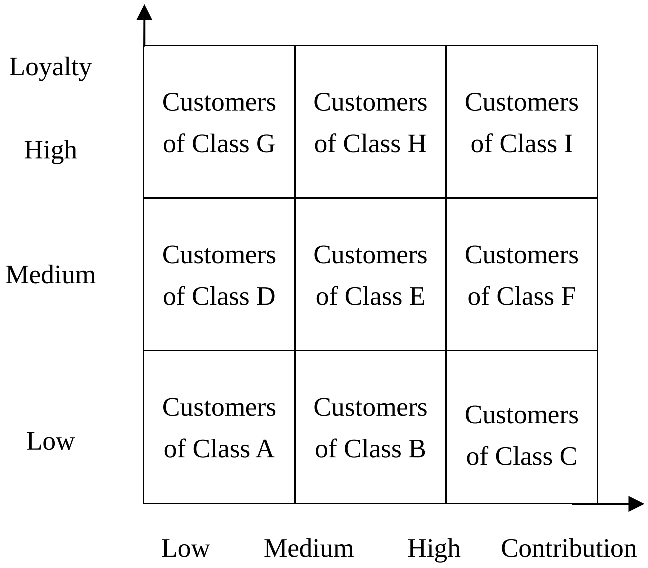
III. APPLICATIONS OF DATA MINING IN BANK'S CRM

A. Application Background

Nowadays, financing products are becoming more alike, and this is a serious problem existing in most of the commercial banks in China. It is of great significance for the performance and development of the banks to provide differentiated marketing strategies aiming at different customers. As a result, it is really necessary for the banks to classify large volume of customers. In this paper, a decision tree is created by using ID3 Algorithm to establish a Classification Model, which can train us to identify the information of the qualified customers who meet the requirements of opening Elite Club Accounts from massive customer data. Furthermore, the high-value customers, the important customers, the ordinary customers of the banks can be identified and their corresponding features can be analyzed in order to provide them with customized services. Hence, their loyalty and contribution to the banks can be further strengthened.

The customers of the banks can be divided into different categories according to their loyalty and contribution to the banks, which is based on their consuming behaviors. Consequently, the customer segmentation model based on customer value can be achieved. Loyalty can be evaluated through the withdrawal histories from the ATM, records of counter transactions in the banks and consuming histories of the POS terminals to see whether or not this customer is using a certain bank as the primary one. Contribution is referred to the profits a customer makes for a certain bank which is mainly embodied in the histories of deposits and loans as well as purchasing records of financing products.

A Customer Value Matrix is then formed as a result of the segmentation, shown in Figure 1.



Loyalty			
High	Customers of Class G	Customers of Class H	Customers of Class I
Medium	Customers of Class D	Customers of Class E	Customers of Class F
Low	Customers of Class A	Customers of Class B	Customers of Class C
	Low	Medium	High

Figure1. Customer Value Matrix

B. Pre-Processing of Data

1) Attribute Selection and Data Protocol

The data used in Data Mining should truly reflect the basic information of the customers and their financial relationships with the banks in order to ensure Data Mining comprehensive, objective, scientific and accurate. This paper makes an analysis of relevant information about individual customers of the commercial banks with the following attributes selected as the inputs for the models: sex, age, education background, income, times of consuming through POS terminals each month, whether purchasing financing products, etc.

Table 1 Table of Customer Attribute Protocol

Attribute	Attribute value	Protocol value
Sex	Male	1
	Female	2
Age	Below 25	1
	25—50	2
	Above 50	3
Education background	High school or lower	1
	College	2
	Master or higher	3
Income(Yuan/month)	Less than 1500	1
	1500—3000	2
	More than 3000	3
Credit records through POS terminals (times/month)	Less than 5	1
	5—10	2
	More than 10	3
Whether purchasing financing products	Yes	1
	No	2

In order to avoid any inconsistency of data, data protocol for massive and complex customer information must be formed in order to store such information in the data sheets. Data protocol can be achieved by sorting data about the customers into certain categories, which can guarantee consistency and comparability of data and ensure that appropriate Data Mining methods and models are applied in for analysis. See Table 1 for more details about the attribute protocol.

2) Algorithm Choice

The methods of creating the customer classification model are classified into decision tree, neural network and clustering methods. The decision tree algorithm is a common data mining algorithm, which is a classification function developed gradually from the realm of machine learning. Its structure is simple, understandable and efficient, which is suitable for a large-scale training set. It can predict the type of customers via analyzing the background information of specific customers by applying the data mining technology of the decision tree to the bank's CRM, and adopt different marketing strategies to improve the service efficiency and boost profits of the bank. The decision tree algorithm includes ID3 algorithm, C4.5 algorithm, CART algorithm, and SLIQ algorithm, among which the ID3 algorithm is the most representative. This paper will use the ID3 algorithm to construct a decision tree for customer classification.

3) Introduction of Algorithm

Suppose S is the training sample set, which contains n categories of samples. These categories can be represented by C= {C_1, C_2, …$C_{n,}$}. Suppose the number of training case of the category i is C and the total number of the training cases is X. If the probability of one case belonging to category i is p_i, then $p_i = \frac{C}{X}$, then the entropy of S is $En(S) = -\sum_{i=0}^{n} p_i \log_2 p_i$.Through this formula it can be viewed that the more uniform the probability distribution of the sample is, the larger its entropy is, and the more miscellaneous the sample set is. Therefore, the entropy can be regarded as a measure of the sample set's impurity level. The larger the entropy is, the higher the impurity level is. The branch principle of decision tree is to make the subsets of the divided sample as pure as possible, namely, to make their entropy as small as possible.

If test attribute A is selected for test, suppose the attribute A has the nature a_1, a_2, a_3…a_j, and if A=a_j, the number of cases that belong to category i is C_{ij}. Record $p(C_i / A = a_j)$ as $p(C_i / A = a_j) = \frac{C_{ij}}{X}$, which is the probability of sample case belonging to category i when the value of the test attribute A equals $a_{j.}$

Suppose attribute A divides S into m segments, the entropy of the subsets divided by A can be represented with the following formula:

$$En(S,A) = \sum_{i=0}^{m} p(C_i / A = a_j) En(S)$$

The information gain is used to measure the reduced value of the entropy, therefore the information gain acquired by using attribute A to divide S is:

$$Gain(S,A) = En(S) - En(S,A)$$

Gain(S, A) is the expected compression of the entropy in case of knowing the value of the attribute A. According to the definition of information gain, the more the entropy is reduced, the larger the information gain is, and the purer the node is, so the larger Gain (S,A) is, the more the information gain of the division volume by selecting the test attribute A. In general, sequence can be made according to the information gain of each attribute. The attribute getting the highest information gain is selected as a branch attribute.

4) Algorithm Implementation

To illustrate the implementation method of this algorithm, this paper gives an example of the training sample set. The customer information provided according to the attribute protocol in table 1 is shown in table 2.

From table 2, it can be viewed that all the five attributes of each sample: sex, age, education, income, and credit record through POS terminal are classification attributes. The class label is represented by the attribute "whether purchasing financing products". The purpose of the ID3 algorithm is to construct a decision tree according to these training samples in order to analyze the roles played by each attribute in "whether purchasing financing products", and whether the customer purchases financing product is the basic condition to become an high-value customer. Therefore, to construct a decision tree is actually a process of searching for high-value customers through classification attributes.

Table 2 Training Sample Set

Customer No.	Sex	Age	Edu	Income	Credit Records through POS terminal	Whether purchasing financial products
1	1	1	2	1	1	1
2	1	1	1	3	2	1
3	1	1	1	1	2	2
4	2	2	2	2	3	1
5	1	1	2	3	2	1
6	2	2	2	3	1	1
7	2	1	1	2	1	2
8	2	2	1	1	2	2
9	1	3	2	2	2	1
10	1	3	3	3	1	1
11	2	3	2	2	2	2
12	2	2	3	1	1	1
13	2	3	2	2	2	2
14	1	2	3	3	3	1
15	1	3	2	2	2	2

In the training sample set S of Table 2, there are nine samples that belong to the first "Whether purchasing financing products" class, and six samples that belong to the second class. The entropy necessary for classifying the given sample is:

$$En(S) = -\frac{9}{15}\log_2\frac{9}{15} - \frac{6}{15}\log_2\frac{6}{15} = 0.971 \cdot$$

The entropy value of 0.971 reflects the uncertainty of the classification of sample set S, which is also the expected information of the sample classification. The following will respectively calculate the information gain acquired by dividing the training samples according to the attributes of income, sex, age, education, and credit record through POS terminal.

Suppose S is the training sample set and income divides S into three parts: income1=1, income2=2, income3=3. If S_v is used to represent the sample set whose attribute value is v, then $S_{income1}$=4, $S_{income2}$=6, $S_{income3}$=5. In $S_{income1}$, there are two samples of Class 1, and two of Class 2. The entropy of $S_{income1}$ is:

$$En(S_{income1}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

Similarly, it is possible to calculate the entropy of $S_{income2}$ and $S_{income3}$ are 0.918 and 0, therefore the entropy of S divided by the attribute income is:

$$En(S, income) = \frac{4}{15}\times 1 + \frac{6}{15}\times 0.918 + \frac{5}{15}\times 0 = 0.634 \quad ,$$

the information gain of income is $Gain(S, income) = 0.971 - 0.634 = 0.337$. Similarly, the other information gain of income is as follows: $Gain\ (S, sex) = 0.079$, $Gain\ (S, age) = 0.083$, $Gain(S, edu) = 0.246$, and $Gain(S, pos) = 0.222$. As the information gain of the attribute income is the highest, the income is selected as the test attribute of root node. The information gain of edu and pos is second to that of income; therefore they can act as the branch nodes of the second and third levels under the root node. The information gain of age and sex is comparatively low, so they can act as branch nodes of lower levels. The model of decision tree is shown in Figure 2 (Due to length of the paper and comparatively low information gain of age and sex, the branch nodes generated by age and sex are deleted in Figure 2).

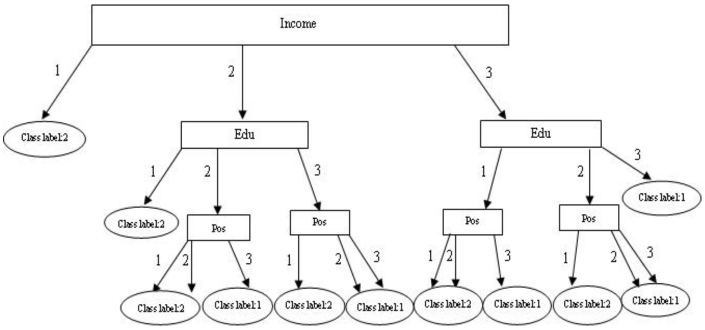


Figure 2 Decision Tree of Customer Classification

After the decision tree is generated, customers can be classified. For example, in the case of attribute income=1, the customer's class label (whether purchasing financial products) is 2; in the case of attribute income=2, attribute edu=2 in the second level and attribute pos=2 in the third level, then his class label is 1. This classification method is not absolute, but it is a reliable way for customer classification.

5) Practical Application of Data Mining Technology

This paper takes the customer data of a service point of Yubei Office, Chongqing Branch of Agricultural Bank of China as the training set. The samples include 53872 customers, who can be classified into three classifications. The data of a service point of Yuzhong Office, Chongqing Branch of Agricultural Bank of China are used as test data for verifying the accuracy of the mode. The test samples include 51749 customers. This research project extracts the basic information of the bank's customers, and various kinds of transaction information between the banks and customers. In addition, data is processed, which is downloaded from the data source files to the ODS library, and then loaded to the central database after classified and combined. Then through the Microsoft Visual Studio tool of SQL Sever2005, the Integration Services project is created to connect multiple data sources, preprocess the data and establish a database. At last, the bank's CRM decision tree model is acquired through the ID3 algorithm.

During the study process of this project, the decision tree that can adapt to the classification management of the bank's customers is constructed through judgment, analysis and adjustment. Accuracy of the prediction this decision tree made reaches 88%. Through this decision tree, some predictive conclusions can be made: 1. The credit quality of the customers plays a vital role in the healthy development of a bank. 2. The preliminary sifting for customer classification comes from monthly income of the customers. 3. The monthly income is not absolute although it is important for a customer to become a high-value one. 4. The education background and credit record through POS terminal of a customer can be regarded as an important reference factor for a customer to become a high-value customer.

IV. Conclusion

According to the attribute features of the bank's customers, this paper combines CRM and the data mining tool, and constructs a decision tree suitable for the bank's customer classification by applying the ID3 algorithm to optimize the methods of classifying the bank's customers. It can be viewed that the introduction of data mining technology into the bank's CRM can realize CRM target with high-quality, bring CRM into full play, effectively improve service quality provided by the bank to target customers and reduce the operation cost for the bank. With the maturity of data mining technology, various kinds of CRM realized on the basis of data mining technology will undoubtedly have a prosperous future.

References

[1] Xun Liang ,Data mining algorithms and application, Peking university press ,April .2006.

[2] Ding Sheng Wan, Chong Liu, Yuan Sheng Liu, Application of Data Mining in the Customer Segmentation of Bank, Microcomputer Applications,Vol 25,pp 31—34,No.2,2009.

[3] Patnaik, Debprakash, Sastry P. S., Unnikrishnan K. P, Inferring neuronal network connectivity from spike data: A temporal data mining approach, Scientific Programming, 2008, Vol. 16 Issue 1, p49-77.

[4] Lu Sun,Jun Yang,Mahmassani, Hani,Wenjun Gu,Bum-Jin Kim, Data mining-based adaptive regression for developing equilibrium speed-density relationships, Canadian Journal of Civil Engineering, Mar2010, Vol. 37 Issue 3, pp389-400,

[5] Guoyin Wang, Yan Wang, 3DM: Domain-oriented Data-driven Data Mining,Fundamenta Informaticae, 2009, Vol. 90 Issue 4, pp395-426.

[6] Thawornwong S,Enke D, The adaptive selection of financial and economic variables for use with artificial neural networks. Neurocomputing, 56:pp205—232,2004.

[7] Jing Ping Zhou, The problems and solution in the application of analyzing CRM to commercial bank based on data mining，Journal of Hubei University(Natural Science),Vol 32,pp46—49,Mar 2010.

[8] Huttenhower, Curtis, Hofmann, Oliver, A Quick Guide to Large-Scale Genomic Data Mining, PLoS Computational Biology, May2010, Vol. 6 Issue 5, pp1-6.

[9] Feng Xiao, Hai Jian Zheng, Chuang Lu, Study on CRM of Bank Based on Clustering Analysis, Technology Economics,Vol 29,PP 87—93,January 2010.

[10] Jian Wang, Hua Ming Lu, Application of data mining technology in CRM, Journal of Beijing Information Science & Technology University,Vol 25,pp84—88, June 2010.

[11] Ruo Wu Zhong, Hui ping Wang, Research on CRM system based on data minin, Journal of Shaoguan University, Vol31, pp29—32, Mar 2010.